Algorithm for Rapid Reconstruction of Protein Backbone from Alpha Carbon Coordinates

MARIUSZ MILIK,[1]*, ANDRZEJ KOLINSKI,[1,2] and JEFFREY SKOLNICK[1]

[1]*The Scripps Research Institute, Department of Molecular Biology, 10666 North Torrey Pines Road, La Jolla, California 92037; and* [2]*Department of Chemistry, University of Warsaw, Pasteura 1, 02093 Warsaw, Poland*

Received 5 September 1995; accepted 4 April 1996

ABSTRACT

A method for generating a full backbone protein structure from the coordinates of α-carbons, is presented. The method extracts information from known protein structures to generate statistical positions for the reconstructed atoms. Tests on a set of proteins structures show the algorithm to be of comparable accuracy to existing procedures. However, the basic advantage of the presented method is its simplicity and speed. In a test run, the present program is shown to be much faster than existing database searching algorithms, and reconstructs about 8000 residues per second. Thus, it may be included as an independent procedure in protein folding algorithms to rapidly generate approximate coordinates of backbone atoms.

Introduction

Many molecular modeling protocols require algorithms for the reconstruction of protein backbone atoms from the positions of Cα carbons. Recently, several algorithms for full-atom backbone reconstruction[1–8] have been published. Basically, two kinds of approaches are used: exploitation of similarity to small fragments of known protein structures and minimization of the local molecular energy. Some of the methods combine both of these approaches.

The method proposed here was prepared to be used ''in flow'' in protein simulation algorithms. The emphasis here was placed upon the minimization of the calculation time. Fortunately, in spite of its simplicity, the method is of comparable accuracy to existing protocols. Calculation of the predicted backbone atom positions requires access to rather small tables based on statistical data and a few arithmetical calculations per residue. Using the proposed method, the frequent switching between a reduced (Cα-based) representation and an

*Author to whom all correspondence should be addressed.

CCC 0192-8651 / 97 / 010080-06

approximate full atom representation becomes computationally tractable.

Description of Method

The method is based on the observation that, for a given set of three consecutive Cα–Cα vectors, the position of the central peptide plate is well defined. This probably reflects the specific interplay of ϕ–ψ angles restraints (see Fig. 1). The idea presented above is not new. It was originally invoked by Purisima and Scheraga[9] and then used for development of a method of reconstruction of protein backbone and side-chain direction by Payne.[1] In the presented implementation, the local conformation of the peptide backbone fragment is represented by a combination of three internal distances: $d_{i-1,i+1}$; $d_{i,i+2}$; and $d_{i-1,i+2}$, where $d_{i,j}$ denotes the distance between Cα carbons in the ith and jth residues. Because most of the amino acids are chiral, the local conformation representation must contain the information about the chirality of the system to distinguish between left- and right-handed structures. Thus, $d_{i,i+3}$ is defined by:

$$d_{i,i+3} = \chi|\mathbf{v}_{i-1} + \mathbf{v}_i + \mathbf{v}_{i+1}| \qquad (1)$$

where $\chi = \text{sign}[(\mathbf{v}_{i-1} \times \mathbf{v}_i) \cdot \mathbf{v}_{i+1}]$, and $\mathbf{v}_i$ denotes a virtual Cα-to-Cα vector from the ith to $i+1$ residue.

Figure 2 presents histograms of $d_{i,i+2}$ and $d_{i,i+3}$ obtained by the statistical analysis of coordinates of 430 known protein structures from the Brookhaven Protein Data Bank (PDB).[10] Most of the values for $d_{i,i+2}$ lie between 4.6 and 7.6 Å, and

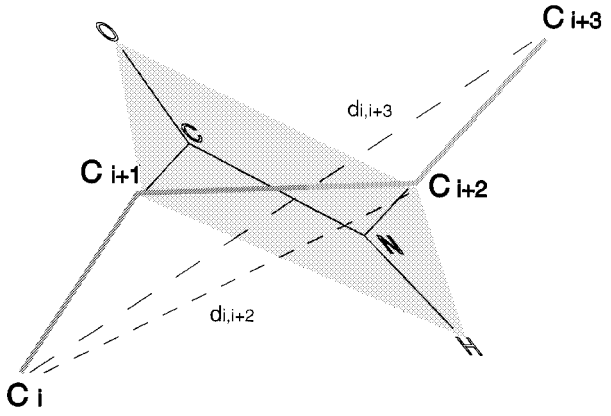

FIGURE 1. Definition of the central peptide plate, $d_{i,i+3}$, and $d_{i,i+2}$. C_i denotes position of the Cα atom of ith residue.

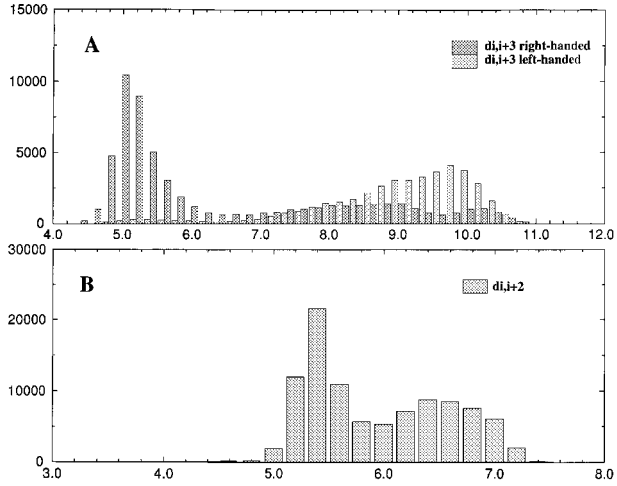

FIGURE 2. Histograms of values of $d_{i,i+3}$ (A) and $d_{i,i+2}$ (B) in 430 protein structures chosen from the Brookhaven Protein Data Bank (see text).

the $d_{i,i+3}$ values lie between 4.2 and 11.0 Å. We clustered the local conformations of four-residue peptide fragments using their values of three internal distances, $d_{i,i+2}$, $d_{i+1,i+3}$, and $d_{i,i+3}$ (defined above). Knowing that the average precision of atomic coordinates in the PDB is about 0.2 Å, we used a 0.3-Å grid in our definition of the local conformation class. This grid gave 10 possible values for $d_{i,i+2}$ and $d_{i+1,i+3}$ and 48 values for $d_{i,i+3}$ (one must remember that the $d_{i,i+3}$ distance is chiral). As a result, we have 4800 possible local conformations for four residue peptide fragments in the presented representation. For every local conformation, we calculated the average positions of C′, O, N, and Cβ atoms in a local coordinate system, defined below.

The three Cα–Cα vectors provide a reference frame in which the coordinates of the peptide bond can be represented. The basic idea of the reference system is shown in Figure 3.

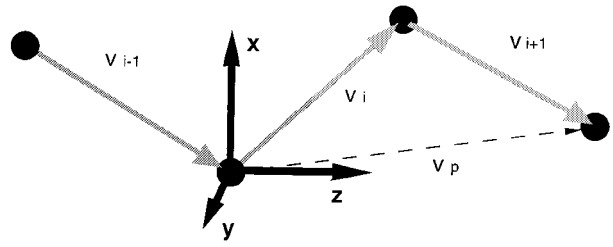

FIGURE 3. Schematic view of the local coordinate system used in the present work.

The first vector in the local coordinate system is constructed according to the formula:

$$\mathbf{x} = (\mathbf{v}_i \times \mathbf{v}_p)/\left(|\mathbf{v}_i \times \mathbf{v}_p|\right) \quad (2)$$

where $\mathbf{v}_p$ is the sum of two consecutive backbone vectors: $\mathbf{v}_p = \mathbf{v}_i + \mathbf{v}_{i+1}$. The second axis is defined as the normalized cross-product of the $\mathbf{v}_p$ and the **x**-axis, defined above:

$$\mathbf{y} = (\mathbf{v}_p \times \mathbf{x})/\left(|\mathbf{v}_p \times \mathbf{x}|\right) \quad (3)$$

and the third axis is orthogonal to **x** and **y**:

$$\mathbf{z} = (\mathbf{x} \times \mathbf{y})/(|\mathbf{x} \times \mathbf{y}|) \quad (4)$$

The local coordination system defined this way depends on the secondary structure of the peptide fragment. An analogous convention for the local coordination system was proposed by Rey and Skolnick.[6]

A table containing average positions of C′, N, O, and Cβ atoms for all types of local peptide structures was stored and used as a starting point in the process of backbone reconstruction. The case of *cis*-proline was treated separately, as it is an amino acid with unusual internal geometry. The data for these residues were stored in a separate database.

The algorithm for all backbone atom reconstruction for a given residue, i, proceeds as follows:

1. Calculate the values of internal distances, $d_{i-1,i+1}$, $d_{i,i+2}$, and $d_{i-1,i+2}$, and chirality (see the definitions in Fig. 1).
2. Find the class of local conformation, according to the values of internal distances and chirality, using the grid defined above.
3. For the given local conformation class, find in the database the average C$\alpha \rightarrow$ C′, C$\alpha \rightarrow$ N, C′ $\rightarrow$ O, and C$\alpha \rightarrow$ Cβ vectors, in the local coordination system. In this case, when the local conformation class is not in our database (this may happen for less ''popular'' states), the algorithm takes the values for the closest represented local conformation class. For *cis*-prolines, these vectors are taken from a separate database, created only for this case. Of course, the C$\alpha \rightarrow$ Cβ vector is not calculated for glycines. In the case of the C$\alpha \rightarrow$ Cβ vector, the statistics are collected for the first Cβ atom in the central peptide plane (the C_{i+1} position in Fig. 1).
4. Build the local coordination system according to Eqs. (1)–(3) (see Fig. 3 and its text description).
5. Rotate the vectors from the local to the laboratory coordinate system and calculate the actual values of atomic coordinates.

Using the above procedure, which starts from Cα coordinates, it is possible to calculate predicted coordinates of all backbone atoms for residues from 2 to $N - 2$ (N is the number of residues in the protein chain under consideration). The problem of calculating the backbone atomic coordinates for the first and two last residues of the chain was approximately solved by using virtual residues on both ends of the chain. The coordinates of the ''dummy'' residues were calculated by repetition of the first (for the beginning of the chain) and last two vectors (for the end).

The proposed algorithm is a step in the application of the homology modeling ideas into the backbone reconstruction problem. The idea of using information about known protein structures in the process of backbone reconstruction is not new (e.g., see Levitt[4]). The basic difference is that our method eliminates the necessity of on-line PDB database searching, and therefore speeds up the reconstruction process without a significant loss in precision. This opens up the possibility of new areas of application such as in the inclusion of the method into Monte Carlo protein structure modeling algorithms.

Results and Discussion

The method was tested on a set of 15 protein structures from the PDB. We have chosen the structures that were used in previous works[1–7] to compare our algorithm with already published methods. The structures used in the process of preparation of our statistical database were excluded from the testing set.

In the testing process, the coordinates of backbone atoms and Cβ were calculated using our algorithm on the basis of Cα coordinates from the PDB file. The predicted coordinates were then compared with the crystallographic ones from PDB, and the distances between the crystallographic and predicted coordinates were calculated. To compare with other methods, we have also calculated root-mean-squared (RMS) distances for these atoms and

the average RMS for backbone atoms. The results are summarized in Table I. For comparison, the same table also presents results obtained for these structures, using the previously published algorithms cited in the present work.[1–7]

Table I presents the values of RMS calculated individually for N, C, O, and Cβ atoms, and the average RMS for all backbone atoms. The comparison with other works shows that, regardless of its speed and simplicity, our algorithm gives a similar level of prediction accuracy as other approaches. The best reconstruction precision is obtained for the nitrogen and carbon atoms, where the RMS is in the range 0.08–0.30 Å. This means that the average error of reconstruction for these atoms is on the level of average error of the experimental data in the PDB. The precision of reconstruction of Cβ atom coordinates is on a similar level with the RMS and is in the range 0.140–0.475 Å. The worst reconstruction precision is obtained from the carbonyl oxygen atoms, whose RMS values range from 0.341 to 0.959 Å; however, the diminished reconstruction accuracy of the carbonyl oxygen atom is typical of other methods as well.[6]

Additional information about the test results are presented in Figure 4 , which shows histograms of errors of reconstruction for the backbone and Cβ atoms for all the molecules from the testing set. Most reconstruction errors for the N and C atoms are less than 0.5 Å, and there is no example where the error is larger than 1.0 Å. A similar situation is obtained for most cases of Cβ atom reconstruction. However, the error is larger than 1.0 Å for several atoms and, in one case, it reached 2.5 Å. The distribution of errors is broader for oxygen atom reconstruction and, in a few cases, errors are larger than 2.0 Å, with a maximal reconstruction error of about 3.5 Å.

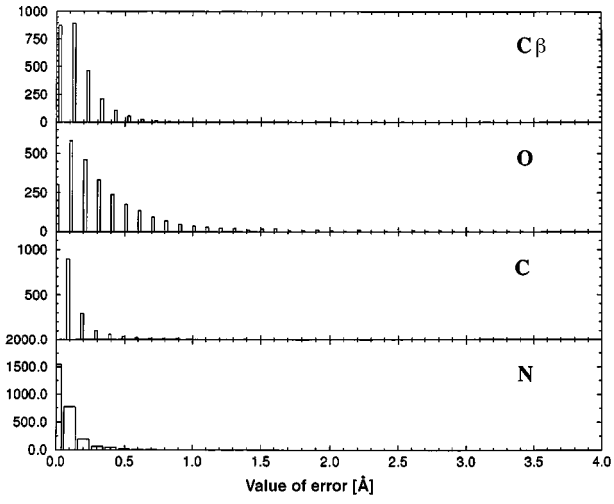

FIGURE 4. Histograms of errors of reconstruction for backbone atoms and Cβ, for the 15 testing proteins, obtained using the present algorithm.

The large values of reconstruction errors occur mostly in the loop or turn fragments and on the surface of proteins. The reconstruction precision is much better for fragments within well-defined secondary structures—this result is typical when a statistical method is used.

Additionally, the quality of the crystallographic structure affects the error of reconstruction. Table II presents errors of reconstruction obtained for different structures of the protein, triosephosphate isomerase (TIM), for chicken (*Gallus gallus*). The worst quality of refinement is obtained for the structure with the poorest resolution (1tim). The precision of backbone reconstruction for other structures is better, and is on the level of the average for the tested set of structures (see Table I).

Figure 5A and B presents individual reconstruction errors for the backbone atoms for testing structures with the best (2wrp, Fig. 5A) and the worst (1tim, Fig. 5B) average RMS.

In the case of the best structure, 2wrp (Fig. 5A), one can see the correlation between errors in reconstruction of nitrogen and carbon atoms and errors in reconstruction of carbonyl oxygens. There is only one position where the error of oxygen reconstruction is larger than 2 Å (residue 60); all other backbone atoms are reconstructed with error on the level of the average experimental error.

Figure 5B presents the worst case, for the 1tim structure. The reconstruction of the nitrogen and carbon atoms is satisfactory. Even in this case, most of the errors are smaller than 0.5 Å. However, the errors of reconstruction of carbonyl oxygens are much larger and, in eight cases, they exceed 2 Å. Most of the large errors happened in residues situated on the surface of the 1tim structure.

In conclusion, the method described here is of comparable accuracy to existing algorithms, but it is much faster. In the test run, on a Sparc 10 work station with a GNU C compiler without optimization, our algorithm reconstructs about 8000 residues per second.

The presented results offer the possibility of several new applications. On the most trivial level,

TABLE I.
Values of RMS Distances for N, C′, O, and Cβ Atoms Calculated for Testing Protein Structures Using the Present Algorithm.[a]

	RMS distances from this work (Å)					RMS distances from prior works (Å)						
Protein	N	C	O	Cβ	Backbone average	Correa[2]	Classens et al.[5]	Holm and Sander[3]	Levitt[4]	Payne[1]	Reid and Thornton[7]	Rey and Skolnick[6]
2alp	0.182	0.243	0.724	0.241	0.453	0.19	x	x	x	0.30	x	x
3app	0.188	0.230	0.657	0.279	0.416	x	x	x	x	x	x	x
5cpa	0.245	0.267	0.749	0.320	0.480	x	0.61	x	x	0.33	x	x
1crn	0.141	0.208	0.660	0.194	0.408	x	x	x	0.56	0.20	x	x
1ctf	0.212	0.265	0.721	0.267	0.461	x	x	x	0.29	0.19	x	x
2cts	0.162	0.197	0.587	0.249	0.369	x	0.54	x	x	0.33	x	x
4fxn	0.164	0.204	0.668	0.223	0.414	x	x	x	x	0.39	x	x
3fxn	0.194	0.270	0.842	0.345	0.522	0.49	x	x	0.44	x	0.51	0.77
2mhr	0.208	0.269	0.714	0.321	0.457	x	x	x	x	0.22	x	0.78
2prk	0.145	0.190	0.672	0.224	0.358	x	x	0.49	x	0.26	x	x
6pti	0.149	0.186	0.616	0.400	0.381	x	x	x	0.51	0.32	x	x
1tim	0.231	0.301	0.959	0.475	0.595	x	0.55	x	x	0.50	x	0.68
1ubq	0.112	0.171	0.523	0.198	0.324	x	x	0.42	x	0.25	x	x
9wga	0.166	0.233	0.725	0.359	0.450	x	x	x	x	x	x	x
2wrp	0.085	0.108	0.341	0.140	0.212	x	x	0.46	x	0.18	x	x

[a]The backbone atom reconstruction results obtained for these structures in previously published studies are presented for comparison.

TABLE II.
Comparison of Reconstruction Errors for Different Structures of Triosephosphate Isomerase (TIM).

Structure	Resolution (Å)	N RMS (Å)	C RMS (Å)	O RMS (Å)	Cβ RMS (Å)	Backbone average (Å)
1tph	1.8	0.124	0.164	0.522	0.189	0.324
1tim	2.5	0.231	0.301	0.959	0.475	0.595
1tpu	1.9	0.125	0.151	0.475	0.172	0.297
1tpb	1.9	0.121	0.156	0.503	0.180	0.312
1tpc	1.9	0.135	0.180	0.574	0.261	0.356
1tpv	1.9	0.129	0.161	0.505	0.174	0.315
1tpw	1.9	0.143	0.186	0.591	0.193	0.367

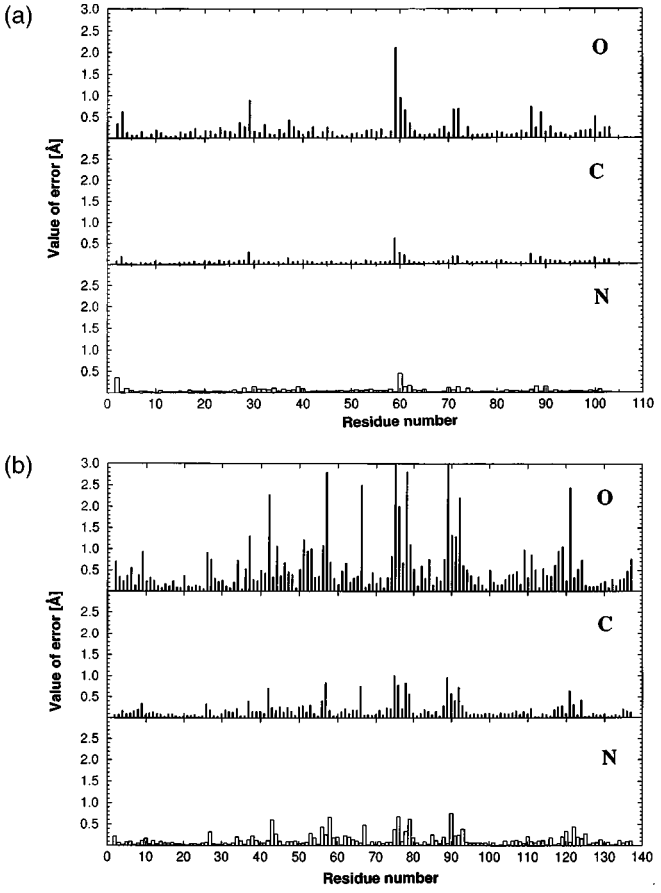

FIGURE 5. Individual reconstruction errors for backbone atoms for 2wrp (A) and the first 140 residues of 1tim (B) structures.

it can be used to generate good starting points for much more expensive and accurate procedures. However, the main area of application is expected to be in folding simulation algorithms employing a reduced Cα representation of protein conformations,[11,12] where the energy evaluation may require approximate coordinates of the backbone atoms.[13] In such algorithms, the backbone reconstruction has to be performed on the order of 10^7 times in a single simulation run. The proposed protocol makes this kind of task feasible.

Acknowledgments

The list of protein structures and statistical data used for these calculations are available via anonymous ftp scripps.edu from directory /pub /milik/backbone. The research was supported by NIH Grant No. GM-37408. A. Kolinski was partly supported by the University of Warsaw Grant #BST 502-34/95. A.K. is an International Research Scholar of the Howard Hughes Medical Institute.

References

1. P. W. Payne, *Prot. Sci.*, **2**, 315–324 (1993).
2. P. E. Correa, *Proteins*, **7**, 366 (1990).
3. L. Holm and C. Sander, *J. Mol. Biol.*, **218**, 183 (1991).
4. M. Levitt, *J. Mol. Biol.*, **226**, 507 (1992).
5. M. Classens, E. van Cutsem, I. Lasters, and S. Wodak, *Prot. Eng.*, **2**, 335 (1989).
6. A. Rey and J. Skolnick, *J. Comput. Chem.*, **13**, 443 (1992).
7. L. Reid and J. Thornton, *J. Prot.*, **5**, 170 (1990).
8. A. Liwo, M. R. Pincus, R. J. Wawak, S. Rackovsky, and H. A. Sheraga, *Prot. Sci.*, **2**, 1697 (1993).
9. E. O. Purisima and H. Scheraga, *Biopolymers*, **23**, 1207 (1984).
10. F. C. Bernstein, et al., *J. Mol. Biol.*, **112**, 535 (1977).
11. A. Kolinski and J. Skolnick, *Proteins*, **18**, 353 (1994).
12. A. Kolinski and J. Skolnick, *Proteins*, **18**, 338 (1994).
13. A. Kolinski, M. Milik, J. Rycombel, and J. Skolnick, *J. Chem. Phys.*, **103**, 4312 (1995).